

Σύντομη περιγραφή

Δεδομένα

Το dataset που έχουμε σκοπό να χρησιμοποιήσουμε προέρχεται από το kaggle.com και αφορά ενέργειες (view, add_to_cart, transaction) πελατών και επιλογές προϊόντων κάποιου online καταστήματος. Τα δεδομένα περιλαμβάνονται σε 3 csv αρχεία και είναι διαθέσιμα [εδώ](#). Εκτός από τις ενέργειες των πελατών, τα δεδομένα περιέχουν πληροφορίες για τα προϊόντα και τις χρονικές στιγμές των γεγονότων.

Ο στόχος είναι να επεξεργαστούμε κατάλληλα τα δεδομένα ώστε να οδηγηθούμε σε κατηγοριοποιήσεις και προβλέψεις μελλοντικής συμπεριφοράς.

Πλάνο υλοποίησης

Οι τεχνολογίες που έχουμε σκοπό να χρησιμοποιήσουμε είναι: Kafka, Spark, Hdfs. Συγκεκριμένα, τα δεδομένα μέσω του Kafka θα αποθηκεύονται σε Hdfs. Στη συνέχεια μέσω του Spark θα πραγματοποιείται η κατηγοριοποίηση και τα αποτελέσματα θα επιστρέφουν στο Hdfs. Το πλάνο υλοποίησης απεικονίζεται στο παρακάτω σχήμα:

